Express Mail Label No. EL474110565US

PATENT

Attorney Docket No. 3334.1

# PATENT APPLICATION

# METHODS AND COMPUTER SOFTWARE PRODUCTS FOR TRANSCRIPTIONAL ANNOTATION


Inventors:

Carsten Rosenow, a citizen of Germany

Residing at 105 Livorno Way,
Redwood City, CA 94065, USA

Thomas Gingeras, a citizen of the United States
Residing at 1541 Crest Dr.
Encinitas, CA 92024

Assignee:

Affymetrix, Inc.
a Corporation Organized under the laws of Delaware


Entity: Large


Affymetrix, Inc.
Attn: Legal Department
3380 Central Expressway
Santa Clara, CA 95051
(408) 731-5000

# METHODS AND COMPUTER SOFTWARE PRODUCTS FOR TRANSCRIPTIONAL ANNOTATION

## RELATED APPLICATION

~~This application claims the priority of U.S. Provisional Application Serial Number 60/205,432, Attorney Docket Number 3334, filed on May 19, 2000, which is incorporated herein by reference.~~

## FIELD OF THE INVENTION

This invention relates to genetic analysis and bioinformatics.

## BACKGROUND OF THE INVENTION

Many biological functions are carried out by regulating the expression levels of various genes, either through changes in the copy number of the genetic DNA, through changes in levels of transcription (*e.g.* through control of initiation, provision of RNA precursors, RNA processing, *etc.*) of particular genes, or through changes in protein synthesis. For example, control of the cell cycle and cell differentiation, as well as diseases, are characterized by the variations in the transcription levels of a group of genes.

The complete genomic sequences of greater than 20 bacterial genomes are already publicly available. However, the transcribed regions of the genomes are largely unknown. Methods for discovering the transcribed regions and their functional significance are needed.

## SUMMARY OF THE INVENTION

The current invention provides methods and computer software products for identifying transcripts. The methods and computer software products have applications in genetic research, drug discovery, diagnostics and pharmacogenetics.

In some embodiments, nucleic acid probes, against a region of a genome, are hybridized with a biological sample derived from the species with the genome. The hybridization signals are analyzed to determine potential transcripts. A region of the genome where the intensity of hybridization of all the probes are above a threshold value (usually the level of non-specific hybridization) is identified. The region may be identified by aligning the probes against the genome; walking through the genome to find regions where all consecutive probes have intensities above the threshold value. In some embodiments, the threshold value may be the intensity of a corresponding mismatch probe.

The regions identified potentially correspond to one transcript. In some embodiments, the intensities of hybridization of probes within each of the identified regions are further analyzed to detect changes in magnitude. For example, in an identified region covered by 40 probes, if the intensities of the first 20 probes have similar intensities, the last 20 probes also have similar intensities; and the intensities of the first 20 probes are twice as much as these of the last 20 probes, the region may contain at least two separately transcribed areas: the first is covered by the first 20 probes and the second is covered by the last 20 probes.

In some embodiments, probe intensities are used to identify locations of transcripts in respect to the computational annotation. Multiple probes with similar intensities give a higher probability of a transcript.

In some other embodiments, probe intensities in non-coding regions are used to identify previously unidentified transcripts. Multiple probes with similar intensities give a higher probability of a transcript.

In some additional embodiments, probe intensities at the 5' and 3' end of a coding region are used to identify potential regulatory regions, termination sequences or regions with unknown function.

In some further embodiments, multiple adjacent transcripts including coding and non-coding regions with similar probe intensities give a high probability of an operon.

In another aspect of the invention, computer software products for transcriptional annotation are provided. In some embodiments, the computer software product contains computer program code for inputting probe intensities, computer program code for identifying genomic regions wherein the intensities of probes are above a threshold value, and code for comparing the intensities of the probes the genomic regions to identify areas where the probe intensities are similar. Each of the areas may be indicated as a potential transcript. The codes may be stored in any computer readable media including a CD-ROM or a hard-drive.

## BRIEF DESCRIPTION OF THE DRAWINGS

The accompanying drawings, which are incorporated in and form a part of this specification, illustrate embodiments of the invention and, together with the description, serve to explain the principles of the invention:

Figure 1 illustrates an example of a computer system that may be utilized to execute the software of an embodiment of the invention.

Figure 2 illustrates a system block diagram of the computer system of Fig. 1.

Figure 3 shows one design of the expression arrays useful for the embodiments of the invention.

Figure 4 shows identification of the start of transcription and transcript extensions.

Figure 5 shows transcript extension.

Figures 6-11 show identification of operons.

Figure 12 shows the identification of regulatory elements.

Figure 13 shows the identification of previously unknown transcript.

## DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

Reference will now be made in detail to the preferred embodiments of the invention. While the invention will be described in conjunction with the preferred embodiments, it will be understood that they are not intended to limit the invention to these embodiments. On the contrary, the invention is intended to cover alternatives, modifications and equivalents, which may be included within the spirit and scope of the invention.

As will be appreciated by one of skill in the art, the present invention may be embodied as a method, data processing system or program products. Accordingly, the present invention may take the form of data analysis systems, methods, analysis software and etc. Software written according to the present invention is to be stored in some form of computer readable medium, such as memory, hard-drive, DVD ROM or CD ROM, or transmitted over a network, and executed by a processor.

Fig. 1 illustrates an example of a computer system that may be used to execute the software of an embodiment of the invention. Fig. 1 shows a computer system 1 that includes a display 3, screen 5, cabinet 7, keyboard 9, and mouse 11. Mouse 11 may have one or more buttons for interacting with a graphic user interface. Cabinet 7 preferably houses a CD-ROM or DVD-ROM drive 13, system memory and a hard drive (*see,* Fig. 2) which may be utilized to store and retrieve software programs incorporating computer code that implements the invention, data for use with the invention and the like. Although a CD 15 is shown as an exemplary computer readable medium, other computer readable storage media including floppy disk, tape, flash memory, system memory, and hard drive may be utilized. Additionally, a data signal embodied in a carrier wave (*e.g.,* in a network including the internet) may be the computer readable storage medium.

Fig. 2 shows a system block diagram of computer system 1 used to execute the software of an embodiment of the invention. As in Fig. 1, computer system 1 includes monitor 3, and keyboard 9, and mouse 11. Computer system 1 further

includes subsystems such as a central processor 51, system memory 53, fixed storage 55 (*e.g.*, hard drive), removable storage 57 (*e.g.*, CD-ROM), display adapter 59, sound card 61, speakers 63, and network interface 65. Other computer systems suitable for use with the invention may include additional or fewer subsystems. For example, another computer system may include more than one processor 51 or a cache memory. Computer systems suitable for use with the invention may also be embedded in a measurement instrument. The embedded systems may control the operation of, for example, a GeneChip® Probe array scanner as well as executing computer codes of the invention.

In one aspect of the invention, methods, computer software products are provided to annotate transcripts of a genome according transcription pattern (the form and quantitative of transcripts). The methods involve the detection of transcription and analysis of the transcription patterns.

## I. TRANSCRIPT DETECTION

### A) NUCLEIC ACID SAMPLES

The transcription pattern (the form and quantitative of transcripts) may be determined by examining a sample containing the transcripts. In some preferred embodiments, a biological sample from cells of the species of interest (the species whose genome is to be annotated, for example, E. coli., yeast, dog or human) is obtained and a nucleic acid sample is prepared.

One of skill in the art will appreciate that it is desirable to have nucleic samples containing target nucleic acid sequences that reflect the transcripts of the

cells of interest. Therefore, suitable nucleic acid samples may contain transcripts of interest. Suitable nucleic acid samples, however, may contain nucleic acids derived from the transcripts of interest. As used herein, a nucleic acid derived from a transcript refers to a nucleic acid for whose synthesis the mRNA transcript or a subsequence thereof has ultimately served as a template. Thus, a cDNA reverse transcribed from a transcript, an RNA transcribed from that cDNA, a DNA amplified from the cDNA, an RNA transcribed from the amplified DNA, etc., are all derived from the transcript and detection of such derived products is indicative of the presence and/or abundance of the original transcript in a sample. Thus, suitable samples include, but are not limited to, transcripts of the gene or genes, cDNA reverse transcribed from the transcript, cRNA transcribed from the cDNA, DNA amplified from the genes, RNA transcribed from amplified DNA, and the like. Transcripts, as used herein, may include, but not limited to pre-mRNA nascent transcript(s), transcript processing intermediates, mature mRNA(s) and degradation products.

In one embodiment, such sample is a homogenate of cells or tissues or other biological samples. Preferably, such sample is a total RNA preparation of a biological sample. More preferably in some embodiments, such a nucleic acid sample is the total mRNA isolated from a biological sample. Those of skill in the art will appreciate that the total mRNA prepared with most methods includes not only the mature mRNA, but also the RNA processing intermediates and nascent pre-mRNA transcripts. For example, total mRNA purified with poly (T) column

contains RNA molecules with poly (A) tails. Those poly A+ RNA molecules could be mature mRNA, RNA processing intermediates, nascent transcripts or degradation intermediates.

Biological samples may be of any biological tissue or fluid or cells. Typical samples include, but are not limited to, sputum, blood, blood cells (e.g., white cells), tissue or fine needle biopsy samples, urine, peritoneal fluid, and pleural fluid, or cells therefrom. Biological samples may also include sections of tissues such as frozen sections taken for histological purposes.

Another typical source of biological samples are cell cultures where gene expression states can be manipulated to explore the relationship among genes.

One of skill in the art would appreciate that it is desirable to inhibit or destroy RNase present in homogenates before homogenates can be used for hybridization. Methods of inhibiting or destroying nucleases are well known in the art. In some preferred embodiments, cells or tissues are homogenized in the presence of chaotropic agents to inhibit nuclease. In some other embodiments, RNase are inhibited or destroyed by heart treatment followed by proteinase treatment.

Methods of isolating total RNA are also well known to those of skill in the art. For example, methods of isolation and purification of nucleic acids are described in detail in Chapter 3 of Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993) and Chapter 3 of

Laboratory Techniques in Biochemistry and Molecular Biology: Hybridization With Nucleic Acid Probes, Part I. Theory and Nucleic Acid Preparation, P. Tijssen, ed. Elsevier, N.Y. (1993)).

In a preferred embodiment, the total RNA is isolated from a given sample using, for example, an acid guanidinium-phenol-chloroform extraction method and polyA$^+$ mRNA is isolated by oligo dT column chromatography or by using (dT)n magnetic beads (see, e.g., Sambrook et al., Molecular Cloning: A Laboratory Manual (2nd ed.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989), or Current Protocols in Molecular Biology, F. Ausubel et al., ed. Greene Publishing and Wiley-Interscience, New York (1987)) .

In one particularly preferred embodiment, total RNA is isolated from mammalian cells using RNeasy Total RNA isolation kit (QIAGEN). If mammalian tissue is used as the source of RNA, a commercial reagent such as TRIzol Reagent (GIBCOL Life Technologies). A second cleanup after the ethanol precipitation step in the TRIzol extraction using Rneasy total RNA isolation kit may be beneficial.

Hot phenol protocol described by Schmitt, et al., (1990) Nucleic Acid Res., 18:3091-3092 is useful for isolating total RNA for yeast cells.

Good quality mRNA may be obtained by, for example, first isolating total RNA and then isolating the mRNA from the total RNA using Oligotex mRNA kit (QIAGEN).

Total RNA from prokaryotes, such as E. coli. Cells, may be obtained by following the protocol for MasterPure complete DNA/RNA purification kit from Epicentre Technologies (Madison, WI).

Frequently, it is desirable to amplify the nucleic acid sample prior to hybridization. One of skill in the art will appreciate that whatever amplification method is used, if a quantitative result is desired, care must be taken to use a method that maintains or controls for the relative frequencies of the amplified nucleic acids to achieve quantitative amplification.

Methods of "quantitative" amplification are well known to those of skill in the art. For example, quantitative PCR involves simultaneously co-amplifying a known quantity of a control sequence using the same primers. This provides an internal standard that may be used to calibrate the PCR reaction. The high density array may then include probes specific to the internal standard for quantification of the amplified nucleic acid.

Other suitable amplification methods include, but are not limited to polymerase chain reaction (PCR) (Innis, et al., PCR Protocols. A guide to Methods and Application. Academic Press, Inc. San Diego, (1990)), ligase chain reaction (LCR) (see Wu and Wallace, Genomics, 4: 560 (1989), Landegren, et al., Science, 241: 1077 (1988) and Barringer, et al., Gene, 89: 117 (1990), transcription amplification (Kwoh, et al., Proc. Natl. Acad. Sci. USA, 86: 1173 (1989)), and self-sustained sequence replication (Guatelli, et al., Proc. Nat. Acad. Sci. USA, 87: 1874 (1990)).

Cell lysates or tissue homogenates often contain a number of inhibitors of polymerase activity. Therefore, RT-PCR typically incorporates preliminary steps to isolate total RNA or mRNA for subsequent use as an amplification template. One tube mRNA capture method may be used to prepare poly(A)+ RNA samples suitable for immediate RT-PCR in the same tube (Boehringer Mannheim). The captured mRNA can be directly subjected to RT-PCR by adding a reverse transcription mix and, subsequently, a PCR mix.

In a particularly preferred embodiment, the sample mRNA is reverse transcribed with a reverse transcriptase and a primer consisting of oligo dT and a sequence encoding the phage T7 promoter to provide a single stranded DNA template. The second DNA strand is polymerized using a DNA polymerase with or without primers (See, U.S. Patent Application Serial Number: 09/102,167, U.S. Provisional Application Serial No. XXX, both incorporated herein by reference for all purposes). After synthesis of double-stranded cDNA, T7 RNA polymerase is added and RNA is transcribed from the cDNA template. Successive rounds of transcription from each single cDNA template results in amplified RNA. Methods of in vitro polymerization are well known to those of skill in the art (see, e.g., Sambrook, supra.) and this particular method is described in detail by Van Gelder, et al., Proc. Natl. Acad. Sci. USA, 87: 1663-1667 (1990). Moreover, Eberwine et al. Proc. Natl. Acad. Sci. USA, 89: 3010-3014 provide a protocol that uses two rounds of amplification via in vitro transcription to achieve greater than $10^6$ fold amplification of the original starting material thereby permitting expression

monitoring even where biological samples are limited. In one preferred embodiment, the in-vitro transcription reaction may be coupled with labeling of the resulting cRNA with biotin using Bioarray high yield RNA transcript labeling kit (Enzo P/N 900182).

Before hybridization, the resulting cRNA may be fragmented. One preferred method for fragmentation employs Rnase free RNA fragmentation buffer (200 mM tris-acetate, pH 8.1, 500 mM potassium acetate, 150 mM magnesium acetate). Approximately 20 μg of cRNA is mixed with 8 μL of the fragmentation buffer. Rnase free water is added to make the volume to 40 μL. The mixture may be incubated at 94 °C for 35 minutes and chilled in ice.

It will be appreciated by one of skill in the art that the direct transcription method described above provides an antisense (aRNA) pool. Where antisense RNA is used as the target nucleic acid, the oligonucleotide probes provided in the array are chosen to be complementary to subsequences of the antisense nucleic acids. Conversely, where the target nucleic acid pool is a pool of sense nucleic acids, the oligonucleotide probes are selected to be complementary to subsequences of the sense nucleic acids. Finally, where the nucleic acid pool is double stranded, the probes may be of either sense as the target nucleic acids include both sense and antisense strands.

The protocols cited above include methods of generating pools of either sense or antisense nucleic acids. Indeed, one approach can be used to generate either sense or antisense nucleic acids as desired. For example, the cDNA can be

directionally cloned into a vector (e.g., Stratagene's p Bluscript II KS (+) phagemid) such that it is flanked by the T3 and T7 promoters. In vitro transcription with the T3 polymerase will produce RNA of one sense (the sense depending on the orientation of the insert), while in vitro transcription with the T7 polymerase will produce RNA having the opposite sense. Other suitable cloning systems include phage lambda vectors designed for Cre-loxP plasmid subcloning (see e.g., Palazzolo et al., Gene, 88: 25-36 (1990)).

The biological sample should contain nucleic acids that reflects the level of at least some of the transcripts present in the cell, tissue or organ of the species of interest. In some embodiments, the biological sample may be prepared from cell, tissue or organs of a particular status. For example, a total RNA preparation from the pituitary of a dog when the dog is pregnant. In another example, samples may be prepared from E. Coli cells after the cells are treated with IPTG. Because certain genes may only be expressed under certain conditions, biological samples derived under various conditions may be needed to observe all transcripts. In some instance, the transcriptional annotation may be specific for a particular physiological, pharmacological or toxicological condition. For example, certain regions of a gene may only be transcribed under specific physiological conditions. Transcript annotation obtained using biological samples from the specific physiological conditions may not be applicable to other physiological conditions.

B) NUCLEIC ACID PROBE ARRAY DESIGN

One preferred method for detection of transcripts uses high density oligonucleotide probe arrays. High density oligonucleotide probe arrays and their use for transcript detection are described in, for example, U.S. Patent Nos. 5,800,992, 6,040,193 and 5,831,070

One of skill in the art will appreciate that an enormous number of array designs are suitable for the practice of this invention. The high density array will typically include a number of probes that specifically hybridize to the sequences of interest including potential and putative transcripts. In addition, in a preferred embodiment, the array will include one or more control probes.

The high density array chip includes test probes. Probes could be oligonucleotides that range from about 5 to about 45 or 5 to about 500 nucleotides, more preferably from about 10 to about 40 nucleotides and most preferably from about 15 to about 40 nucleotides in length. In other particularly preferred embodiments the probes are 20 or 25 nucleotides in length. In another preferred embodiment, test probes are double or single strand DNA sequences. DNA sequences are isolated or cloned from nature sources or amplified from nature sources using nature nucleic acid as templates. These probes have sequences complementary to particular subsequences of the genes whose expression they are designed to detect. Thus, the test probes are capable of specifically hybridizing to the target nucleic acid they are to detect.

In addition to test probes that bind the target nucleic acid(s) of interest, the high density array can contain a number of control probes. The control probes fall into three categories referred to herein as 1) Normalization controls; 2) Expression level controls; and 3) Mismatch controls which are designed to contain at least one base that is different from that of a target sequence. Normalization controls are oligonucleotide or other nucleic acid probes that are complementary to labeled reference oligonucleotides or other nucleic acid sequences that are added to the nucleic acid sample. The signals obtained from the normalization controls after hybridization provide a control for variations in hybridization conditions, label intensity, "reading" efficiency and other factors that may cause the signal of a perfect hybridization to vary between arrays. In a preferred embodiment, signals (e.g., fluorescence intensity) read from all other probes in the array are divided by the signal (e.g., fluorescence intensity) from the control probes thereby normalizing the measurements.

Virtually any probe may serve as a normalization control. However, it is recognized that hybridization efficiency varies with base composition and probe length. Preferred normalization probes are selected to reflect the average length of the other probes present in the array, however, they can be selected to cover a range of lengths. The normalization control(s) can also be selected to reflect the (average) base composition of the other probes in the array, however in a preferred embodiment, only one or a few normalization probes are used and they

are selected such that they hybridize well (i.e. no secondary structure) and do not match any target-specific probes.

Expression level controls are probes that hybridize specifically with constitutively expressed genes in the biological sample. Virtually any constitutively expressed gene provides a suitable target for expression level controls. Typically expression level control probes have sequences complementary to subsequences of constitutively expressed "housekeeping genes" including, but not limited to the ß-actin gene, the transferrin receptor gene, the GAPDH gene, and the like.

Mismatch controls may also be provided for the probes to the target genes, for expression level controls or for normalization controls. Mismatch controls are oligonucleotide probes or other nucleic acid probes designed to be identical to their corresponding test, target or control probes except for the presence of one or more mismatched bases. A mismatched base is a base selected so that it is not complementary to the corresponding base in the target sequence to which the probe would otherwise specifically hybridize. One or more mismatches are selected such that under appropriate hybridization conditions (e.g. stringent conditions) the test or control probe would be expected to hybridize with its target sequence, but the mismatch probe would not hybridize (or would hybridize to a significantly lesser extent). Preferred mismatch probes contain a central mismatch. Thus, for example, where a probe is a 20 mer, a corresponding mismatch probe will have the

identical sequence except for a single base mismatch (e.g., substituting a G, a C or a T for an A) at any of positions 6 through 14 (the central mismatch).

Mismatch probes thus provide a control for non-specific binding or cross-hybridization to a nucleic acid in the sample other than the target to which the probe is directed.

The difference in intensity between the perfect match and the mismatch probe (I(PM)-I(MM)) provides a good measure of the concentration of the hybridized material.

The high density array may also include sample preparation/amplification control probes. These are probes that are complementary to subsequences of control genes selected because they do not normally occur in the nucleic acids of the particular biological sample being assayed. Suitable sample preparation/amplification control probes include, for example, probes to bacterial genes (e.g., Bio B) where the sample in question is a biological from a eukaryote.

The RNA sample is then spiked with a known amount of the nucleic acid to which the sample preparation/amplification control probe is directed before processing. Quantification of the hybridization of the sample preparation/amplification control probe then provides a measure of alteration in the abundance of the nucleic acids caused by processing steps (e.g. PCR, reverse transcription, in vitro transcription, etc.).

In a preferred embodiment, oligonucleotide probes in the high density array are selected to bind specifically to the nucleic acid target to which they are directed with minimal non-specific binding or cross-hybridization under the particular hybridization conditions utilized. Because the high density arrays of this invention can contain in excess of 1,000,000 different probes, it is possible to provide every probe of a characteristic length that binds to a particular nucleic acid sequence. Thus, for example, the high density array can contain every possible 20 mer sequence complementary to an IL-2 mRNA.

There, however, may exist 20 mer subsequences that are not unique to the IL-2 mRNA. Probes directed to these subsequences are expected to cross hybridize with occurrences of their complementary sequence in other regions of the sample genome. Similarly, other probes simply may not hybridize effectively under the hybridization conditions (e.g., due to secondary structure, or interactions with the substrate or other probes). Thus, in a preferred embodiment, the probes that show such poor specificity or hybridization efficiency are identified and may not be included either in the high density array itself (e.g., during fabrication of the array) or in the post-hybridization data analysis.

Probes as short as 15, 20, or 25 nucleotide are sufficient to hybridize to a subsequence of a gene and that, for most genes, there is a set of probes that performs well across a wide range of target nucleic acid concentrations. In a preferred embodiment, it is desirable to choose a preferred or "optimum" subset of probes for each gene before synthesizing the high density array.

In one aspect of the invention, probe arrays for transcript annotation may contain probes that are selected to detect interested regions of a genome. The probes may tile the entire region of interest or select areas of the region of interest. Figure 3 shows the design of one such probe array. Multiple probes are selected to tile the gene sequence of interest without overlapping. Tiling methods and strategies are discussed in substantial detail in Published PCT Application No. 95/11995, the complete disclosure of which is incorporated herein by reference in its entirety for all purposes. In some embodiments, however, overlapping probes may be preferred.

B) FORMING NUCLEIC ACID PROBE ARRAYS

Methods of forming high density arrays of oligonucleotides, peptides and other polymer sequences with a minimal number of synthetic steps are disclosed in, for example, 5,143,854, 5,252,743, 5,384,261, 5,405,783, 5,424,186, 5,429,807, 5,445,943, 5,510,270, 5,677,195, 5,571,639, 6,040,138, all incorporated herein by reference for all purposes. The oligonucleotide analogue array can be synthesized on a solid substrate by a variety of methods, including, but not limited to, light-directed chemical coupling, and mechanically directed coupling. See Pirrung et al., U.S. Patent No. 5,143,854 (see also PCT Application No. WO 90/15070) and Fodor et al., PCT Publication Nos. WO 92/10092 and WO 93/09668 and U.S. Pat. No. 5,677,195 which disclose methods of forming vast arrays of peptides, oligonucleotides and other molecules using, for example, light-directed synthesis techniques. See also, Fodor et al., Science, 251, 767-77 (1991). These procedures

for synthesis of polymer arrays are now referred to as VLSIPS™ procedures. Using the VLSIPS™ approach, one heterogeneous array of polymers is converted, through simultaneous coupling at a number of reaction sites, into a different heterogeneous array. See, U.S. Patent Nos. 5,384,261 and 5,677,195.

The development of VLSIPS™ technology as described in the above-noted U.S. Patent No. 5,143,854 and PCT patent publication Nos. WO 90/15070 and 92/10092, is considered pioneering technology in the fields of combinatorial synthesis and screening of combinatorial libraries.

In brief, the light-directed combinatorial synthesis of oligonucleotide arrays on a glass surface proceeds using automated phosphoramidite chemistry and chip masking techniques. In one specific implementation, a glass surface is derivatized with a silane reagent containing a functional group, e.g., a hydroxyl or amine group blocked by a photolabile protecting group. Photolysis through a photolithogaphic mask is used selectively to expose functional groups which are then ready to react with incoming 5'-photoprotected nucleoside phosphoramidites. The phosphoramidites react only with those sites which are illuminated (and thus exposed by removal of the photolabile blocking group). Thus, the phosphoramidites only add to those areas selectively exposed from the preceding step. These steps are repeated until the desired array of sequences have been synthesized on the solid surface. Combinatorial synthesis of different oligonucleotide analogues at different locations on the array is determined by the

pattern of illumination during synthesis and the order of addition of coupling reagents.

In the event that an oligonucleotide analogue with a polyamide backbone is used in the VLSIPS™ procedure, it is generally inappropriate to use phosphoramidite chemistry to perform the synthetic steps, since the monomers do not attach to one another via a phosphate linkage. Instead, peptide synthetic methods are substituted. See, e.g., Pirrung et al. U.S. Pat. No. 5,143,854.

Peptide nucleic acids are commercially available from, e.g., Biosearch, Inc. (Bedford, MA) which comprise a polyamide backbone and the bases found in naturally occurring nucleosides. Peptide nucleic acids are capable of binding to nucleic acids with high specificity, and are considered "oligonucleotide analogues" for purposes of this disclosure.

In addition to the foregoing, additional methods which can be used to generate an array of oligonucleotides on a single substrate are described in in PCT Publication No. WO 93/09668. In the methods disclosed in the application, reagents are delivered to the substrate by either (1) flowing within a channel defined on predefined regions or (2) "spotting" on predefined regions or (3) through the use of photoresist. However, other approaches, as well as combinations of spotting and flowing, may be employed. In each instance, certain activated regions of the substrate are mechanically separated from other regions when the monomer solutions are delivered to the various reaction sites.

A typical "flow channel" method applied to the compounds and libraries of the present invention can generally be described as follows. Diverse polymer sequences are synthesized at selected regions of a substrate or solid support by forming flow channels on a surface of the substrate through which appropriate reagents flow or in which appropriate reagents are placed. For example, assume a monomer "A" is to be bound to the substrate in a first group of selected regions. If necessary, all or part of the surface of the substrate in all or a part of the selected regions is activated for binding by, for example, flowing appropriate reagents through all or some of the channels, or by washing the entire substrate with appropriate reagents. After placement of a channel block on the surface of the substrate, a reagent having the monomer A flows through or is placed in all or some of the channel(s). The channels provide fluid contact to the first selected regions, thereby binding the monomer A on the substrate directly or indirectly (via a spacer) in the first selected regions.

Thereafter, a monomer B is coupled to second selected regions, some of which may be included among the first selected regions. The second selected regions will be in fluid contact with a second flow channel(s) through translation, rotation, or replacement of the channel block on the surface of the substrate; through opening or closing a selected valve; or through deposition of a layer of chemical or photoresist. If necessary, a step is performed for activating at least the second regions. Thereafter, the monomer B is flowed through or placed in the second flow channel(s), binding monomer B at the second selected locations. In

this particular example, the resulting sequences bound to the substrate at this stage of processing will be, for example, A, B, and AB. The process is repeated to form a vast array of sequences of desired length at known locations on the substrate.

After the substrate is activated, monomer A can be flowed through some of the channels, monomer B can be flowed through other channels, monomer C can be flowed through still other channels, etc. In this manner, many or all of the reaction regions are reacted with a monomer before the channel block must be moved or the substrate must be washed and/or reactivated. By making use of many or all of the available reaction regions simultaneously, the number of washing and activation steps can be minimized.

One of skill in the art will recognize that there are alternative methods of forming channels or otherwise protecting a portion of the surface of the substrate. For example, according to some embodiments, a protective coating such as a hydrophilic or hydrophobic coating (depending upon the nature of the solvent) is utilized over portions of the substrate to be protected, sometimes in combination with materials that facilitate wetting by the reactant solution in other regions. In this manner, the flowing solutions are further prevented from passing outside of their designated flow paths.

High density nucleic acid arrays can be fabricated by depositing presynthezied or nature nucleic acids in predefined positions. As disclosed in the U.S. Patent No. 5,040,138, and its parent applications, previously incorporated by reference for all purposes, synthesized or nature nucleic acids are deposited on

specific locations of a substrate by light directed targeting and oligonucleotide directed targeting. Nucleic acids can also be directed to specific locations in much the same manner as the flow channel methods. For example, a nucleic acid A can be delivered to and coupled with a first group of reaction regions which have been appropriately activated. Thereafter, a nucleic acid B can be delivered to and reacted with a second group of activated reaction regions. Nucleic acids are deposited in selected regions. Another embodiment uses a dispenser that moves from region to region to deposit nucleic acids in specific spots. Typical dispensers include a micropipette or capillary pin to deliver nucleic acid to the substrate and a robotic system to control the position of the micropipette with respect to the substrate. In other embodiments, the dispenser includes a series of tubes, a manifold, an array of pipettes or capillary pins, or the like so that various reagents can be delivered to the reaction regions simultaneously.

C) HYBRIDIZATION OF NUCLEIC ACID SAMPLES TO PROBE ARRAYS

Nucleic acid hybridization simply involves contacting a probe and target nucleic acid under conditions where the probe and its complementary target can form stable hybrid duplexes through complementary base pairing. The nucleic acids that do not form hybrid duplexes are then washed away leaving the hybridized nucleic acids to be detected, typically through detection of an attached detectable label. It is generally recognized that nucleic acids are denatured by increasing the temperature or decreasing the salt concentration of the buffer containing the nucleic acids. Under low stringency conditions (e.g., low

temperature and/or high salt) hybrid duplexes (e.g., DNA:DNA, RNA:RNA, or RNA:DNA) will form even where the annealed sequences are not perfectly complementary. Thus specificity of hybridization is reduced at lower stringency. Conversely, at higher stringency (e.g., higher temperature or lower salt) successful hybridization requires fewer mismatches.

One of skill in the art will appreciate that hybridization conditions may be selected to provide any degree of stringency. In a preferred embodiment, hybridization is performed at low stringency in this case in 6X SSPE-T at 37 C (0.005% Triton X-100) to ensure hybridization and then subsequent washes are performed at higher stringency (e.g., 1 X SSPE-T at 37 C) to eliminate mismatched hybrid duplexes. Successive washes may be performed at increasingly higher stringency (e.g., down to as low as 0.25 X SSPE-T at 37 C to 50 C) until a desired level of hybridization specificity is obtained. Stringency can also be increased by addition of agents such as formamide. Hybridization specificity may be evaluated by comparison of hybridization to the test probes with hybridization to the various controls that can be present (e.g., expression level control, normalization control, mismatch controls, etc.).

In general, there is a tradeoff between hybridization specificity (stringency) and signal intensity. Thus, in a preferred embodiment, the wash is performed at the highest stringency that produces consistent results and that provides a signal intensity greater than approximately 10% of the background intensity. Thus, in a preferred embodiment, the hybridized array may be washed at successively higher stringency solutions and read between each wash. Analysis of the data sets thus produced will reveal a wash stringency above which the hybridization pattern is not appreciably altered and which provides adequate signal for the particular oligonucleotide probes of interest.

In a preferred embodiment, background signal is reduced by the use of a detergent (e.g., C-TAB) or a blocking reagent (e.g., sperm DNA, cot-1 DNA, etc.) during the hybridization to reduce non-specific binding. In a particularly preferred embodiment, the

hybridization is performed in the presence of about 0.5 mg/ml DNA (e.g., herring sperm DNA). The use of blocking agents in hybridization is well known to those of skill in the art (see, e.g., Chapter 8 in P. Tijssen, supra.)

The stability of duplexes formed between RNAs or DNAs are generally in the order of RNA:RNA > RNA:DNA > DNA:DNA, in solution. Long probes have better duplex stability with a target, but poorer mismatch discrimination than shorter probes (mismatch discrimination refers to the measured hybridization signal ratio between a perfect match probe and a single base mismatch probe). Shorter probes (e.g., 8-mers) discriminate mismatches very well, but the overall duplex stability is low.

Altering the thermal stability ($T_m$) of the duplex formed between the target and the probe using, e.g., known oligonucleotide analogues allows for optimization of duplex stability and mismatch discrimination. One useful aspect of altering the $T_m$ arises from the fact that adenine-thymine (A-T) duplexes have a lower $T_m$ than guanine-cytosine (G-C) duplexes, due in part to the fact that the A-T duplexes have 2 hydrogen bonds per base-pair, while the G-C duplexes have 3 hydrogen bonds per base pair. In heterogeneous oligonucleotide arrays in which there is a non-uniform distribution of bases, it is not generally possible to optimize hybridization for each oligonucleotide probe simultaneously. Thus, in some embodiments, it is desirable to selectively destabilize G-C duplexes and/or to increase the stability of A-T duplexes. This can be accomplished, e.g., by substituting guanine residues in the probes of an array which form G-C duplexes with hypoxanthine, or by substituting adenine residues in probes which form A-T duplexes

with 2,6 diaminopurine or by using the salt tetramethyl ammonium chloride (TMACl) in place of NaCl.

Altered duplex stability conferred by using oligonucleotide analogue probes can be ascertained by following, e.g., fluorescence signal intensity of oligonucleotide analogue arrays hybridized with a target oligonucleotide over time. The data allow optimization of specific hybridization conditions at, e.g., room temperature (for simplified diagnostic applications in the future).

Another way of verifying altered duplex stability is by following the signal intensity generated upon hybridization with time. Previous experiments using DNA targets and DNA chips have shown that signal intensity increases with time, and that the more stable duplexes generate higher signal intensities faster than less stable duplexes. The signals reach a plateau or "saturate" after a certain amount of time due to all of the binding sites becoming occupied. These data allow for optimization of hybridization, and determination of the best conditions at a specified temperature.

Methods of optimizing hybridization conditions are well known to those of skill in the art (see, e.g., Laboratory Techniques in Biochemistry and Molecular Biology, Vol. 24: Hybridization With Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., (1993)).

D) SIGNAL DETECTION

In a preferred embodiment, the hybridized nucleic acids are detected by detecting one or more labels attached to the sample nucleic acids. The labels may be incorporated by any of a number of means well known to those of skill in the art. However, in a preferred embodiment, the label is simultaneously incorporated during the amplification

step in the preparation of the sample nucleic acids. Thus, for example, polymerase chain reaction (PCR) with labeled primers or labeled nucleotides will provide a labeled amplification product. In a preferred embodiment, transcription amplification, as described above, using a labeled nucleotide (e.g. fluorescein-labeled UTP and/or CTP) incorporates a label into the transcribed nucleic acids. Alternatively, cDNAs synthesized using a RNA sample as a template, cRNAs are synthesized using the cDNAs as templates using in vitro transcription (IVT). A biotin label may be incorporated during the IVT reaction (Enzo Bioarray high yield labeling kit).

Alternatively, a label may be added directly to the original nucleic acid sample (e.g., mRNA, polyA mRNA, cDNA, etc.) or to the amplification product after the amplification is completed. Means of attaching labels to nucleic acids are well known to those of skill in the art and include, for example nick translation or end-labeling (e.g. with a labeled RNA) by kinasing of the nucleic acid and subsequent attachment (ligation) of a nucleic acid linker joining the sample nucleic acid to a label (e.g., a fluorophore).

Detectable labels suitable for use in the present invention include any composition detectable by spectroscopic, photochemical, biochemical, immunochemical, electrical, optical or chemical means. Useful labels in the present invention include biotin for staining with labeled streptavidin conjugate, magnetic beads (e.g., Dynabeads$^{TM}$), fluorescent dyes (e.g., fluorescein, texas red, rhodamine, green fluorescent protein, and the like), radiolabels (e.g., $^{3}H$, $^{125}I$, $^{35}S$, $^{14}C$, or $^{32}P$), enzymes (e.g., horse radish peroxidase, alkaline phosphatase and others commonly used in an ELISA), and colorimetric labels such as colloidal gold or colored glass or plastic (e.g., polystyrene,

polypropylene, latex, etc.) beads. Patents teaching the use of such labels include U.S. Patent Nos. 3,817,837; 3,850,752; 3,939,350; 3,996,345; 4,277,437; 4,275,149; and 4,366,241.

Means of detecting such labels are well known to those of skill in the art. Thus, for example, radiolabels may be detected using photographic film or scintillation counters, fluorescent markers may be detected using a photodetector to detect emitted light. Enzymatic labels are typically detected by providing the enzyme with a substrate and detecting the reaction product produced by the action of the enzyme on the substrate, and colorimetric labels are detected by simply visualizing the colored label. One particularly preferred method uses colloidal gold label that can be detected by measuring scattered light.

The label may be added to the target (sample) nucleic acid(s) prior to, or after the hybridization. So called "direct labels" are detectable labels that are directly attached to or incorporated into the target (sample) nucleic acid prior to hybridization. In contrast, so called "indirect labels" are joined to the hybrid duplex after hybridization. Often, the indirect label is attached to a binding moiety that has been attached to the target nucleic acid prior to the hybridization. Thus, for example, the target nucleic acid may be biotinylated before the hybridization. After hybridization, an aviden-conjugated fluorophore will bind the biotin bearing hybrid duplexes providing a label that is easily detected. For a detailed review of methods of labeling nucleic acids and detecting labeled hybridized nucleic acids see Laboratory Techniques in Biochemistry and Molecular

Biology, Vol. 24: Hybridization With Nucleic Acid Probes, P. Tijssen, ed. Elsevier, N.Y., (1993)).

Fluorescent labels are preferred and easily added during an in vitro transcription reaction. In a preferred embodiment, fluorescein labeled UTP and CTP are incorporated into the RNA produced in an in vitro transcription reaction as described above.

Means of detecting labeled target (sample) nucleic acids hybridized to the probes of the high density array are known to those of skill in the art. Thus, for example, where a colorimetric label is used, simple visualization of the label is sufficient. Where a radioactive labeled probe is used, detection of the radiation (e.g. with photographic film or a solid state detector) is sufficient.

In a preferred embodiment, however, the target nucleic acids are labeled with a fluorescent label and the localization of the label on the probe array is accomplished with fluorescent microscopy. The hybridized array is excited with a light source at the excitation wavelength of the particular fluorescent label and the resulting fluorescence at the emission wavelength is detected. In a particularly preferred embodiment, the excitation light source is a laser appropriate for the excitation of the fluorescent label.

The confocal microscope may be automated with a computer-controlled stage to automatically scan the entire high density array. Similarly, the microscope may be equipped with a phototransducer (e.g., a photomultiplier, a solid state array, a CCD camera, etc.) attached to an automated data acquisition system to automatically record the fluorescence signal produced by hybridization to each oligonucleotide probe on the array. Such automated systems are described at length in U.S. Patent No: 5,143,854, PCT

Application 20 92/10092, and U.S. Application Ser. No. 08/195,889 filed on February 10, 1994. Use of laser illumination in conjunction with automated confocal microscopy for signal detection permits detection at a resolution of better than about 100 $\mu$m, more preferably better than about 50 $\mu$m, and most preferably better than about 25 $\mu$m.

One of skill in the art will appreciate that methods for evaluating the hybridization results vary with the nature of the specific probe nucleic acids used as well as the controls provided. In the simplest embodiment, simple quantification of the fluorescence intensity for each probe is determined. This is accomplished simply by measuring probe signal strength at each location (representing a different probe) on the high density array (*e.g.*, where the label is a fluorescent label, detection of the amount of florescence (intensity) produced by a fixed excitation illumination at each location on the array). Comparison of the absolute intensities of an array hybridized to nucleic acids from a "test" sample with intensities produced by a "control" sample provides a measure of the relative expression of the nucleic acids that hybridize to each of the probes.

One of skill in the art, however, will appreciate that hybridization signals will vary in strength with efficiency of hybridization, the amount of label on the sample nucleic acid and the amount of the particular nucleic acid in the sample. Typically nucleic acids present at very low levels (*e.g.*, < 1pM) will show a very weak signal. At some low level of concentration, the signal becomes virtually indistinguishable from the background. In evaluating the hybridization data, a threshold intensity value may be selected below which a signal is not counted as being essentially indistinguishable from the background.

E) TRANSCRIPTIONAL ANNOTATION

In some embodiments, nucleic acid probes designed to be complementary to a region of a genome are hybridized with a nucleic acid sample derived from the species with the genome. The hybridization signals are analyzed to determine potential transcripts. A region of the genome where the intensity of hybridization of all the probes are above a threshold value (usually the level of non-specific hybridization) is identified. The region may be identified by aligning the probes against the genome; walking through the genome to find regions where all consecutive probes have intensities above the threshold value. In some embodiments, the threshold value may be the intensity of a corresponding probe that is designed to contain a mismatch.

The regions identified potentially correspond to one or more transcripts. In some embodiments, the intensities of hybridization of probes within each of the identified regions are further analyzed to detect changes in magnitude. For example, in an identified region covered by 40 probes, if the intensities of the first 20 probes have similar intensities, the last 20 probes also have similar intensities; and the intensities of the first 20 probes are twice as much as those of the last 20 probes, the region may contain at least two separately transcribed areas: the first is covered by the first 20 probes and the second is covered by the last 20 probes.

In some embodiments, probe intensities are used to identify locations of transcripts in respect to the computational annotation. Multiple probes with similar intensities give a higher probability of a single transcript.

In some other embodiments, probe intensities in non-coding regions are used to identify previously unidentified transcripts. Multiple probes with similar intensities give a higher probability of a transcript.

In some additional embodiments, probe intensities at the 5' and 3' end of a coding region are used to identify potential regulatory regions, termination sequences or regions with unknown function.

In some further embodiments, multiple adjacent transcripts including coding and non-coding regions with similar probe intensities give a high probability of an operon.

In another aspect of the invention, computer software products for transcriptional annotation are provided. In some embodiments, the computer software product contains computer program code for inputting probe intensities, computer program code for identifying genomic regions wherein the intensities of probes are above a threshold value, and code for comparing the intensities of the probes with the genomic regions to identify areas where the probe intensities are similar. Each of the areas may be indicated as a potential transcript. The codes may be stored in any computer readable media including a CD-ROM or a hard-drive.

## II. Example

This example illustrates the transcriptional annotation of several operons in E. coli.

A) RNA isolation and cDNA synthesis

A single colony of E. coli K-12 (MG1655) was inoculated in 5ml of Luria-Bertini (LB) broth and grown over-night at 37°C. The next day 20ml LB-broth was inoculated

with 0.2ml of the overnight culture and grown at 37°C with constant aeration to an optical density (OD600) of 0.8. IPTG was added 30 minutes before the bacteria were harvested to induce the lac-operon.

Total RNA was isolated from the cells using the MasterPure complete DNA/RNA purification kit from Epicentre Technologies (Madison, WI). Cultures were briefly centrifuged at 5000xg, resuspended in lysis buffer and the RNA isolation proceeded following the manufacture's protocol. Isolated RNA was resuspended in diethyl pyrocarbonate (DEPC)-treated water, quantitated based on the absorption at 260nm and stored in aliquotes at -20°C until further use.

cDNA was synthesized using full length total RNA. 9mg of total RNA was reverse transcribed using random hexamers as primers. The RNA was removed using a RNAseH and RNAseA cocktail. After purification of the cDNA (50% yield) a partial DNAseI digest was performed. The cDNA fragments are biotin labeled with terminal transferase using Biotin-ddATP as the substrate. 2 $\mu$g of the labeled cDNA was hybridized to probe arrays. 70ng full length cDNA was used for operon verification by RT-PCR. Standard PCR was performed using different combinations of forward and reverse primers from each gene. The location of the primers is illustrated in the figures. Templates generated without reverse transcriptase were used as controls.

B). Identification of Start of Transcription and Transcript Extension

Figure 4 shows a genomic region containing genes metA, aceB, aceA and aceK (top boxes). The genes were previously identified by Blattner et al., F. R. Blattner et al. 1997. The complete Genome sequence of Escherichia coli K-12. Science: 277; 1453

- 1462. The vertical bars under the schematic showing the genomic region show intensities of probes. A scale is provided on the left. The probe intensities changes from around 100 to 10,000 in the region between metA and aceB, which indicates that the gene metA is in a different transcript from where gene aceB is. To verify that genes metA and aceB are not transcribed as one unit, a RT-PCR is performed. If the genes metA and aceB are transcribed as one unit, the product of RT-PCR would be a 1.5 kb fragment. As shown in the lower portion of Figure 4, the predicted 1.5 kb fragment was not found, which is consistent with the finding that genes metA and aceB are not transcribed as one unit.

The probes tiling the intergenic regions between metA and aceB were also hybridized with intensities similar to the probes tiling metA and aceB, respectively, which indicates that some of the intergenic regions are also transcribed. Therefore, the transcripts may extend beyond the 5' or 3' of a Blattner gene.

Similarly, in Figure 5, transcripts extending into the intergenic regions between the yadQ and yadR genes were detected.

Figure 6 shows the identification of an operon which contains at least the genes ribF, ileS, lspA, slpA and lytB. The intensities of the probes targeting the intergenic regions between ribF and ileS were similar to these of the probes targeting the ribF and ileS genes. Similarly, the probes targeting the intergenic regions between genes lspA and sipA had similar intensities to probes targeting the genes. The result indicates that the five neighboring genes may be transcribed as an operon. To verify this result, RT-PCRs with predicted products spanning over the intragenic regions were performed. In the

lower portion of the Figure 6 shows the result of the RT-PCR experiments. Transcripts containing the intergenic regions between ribF and ileS gene were identified (0.94 kb fragment). Transcripts containing the intergenic regions between lspA and sipA were also found (0.94 and 1.84 kb fragments). These transcripts verified that the intergenic regions were transcribed as the probe array experiments show.

Similarly, Figures 7-11 show the identification of additional operons using oligonucleotide probe arrays. Some of the results were verified by RT-PCR experiments as described for the operon shown in Figure 6.

Figure 12 shows a genomic region containing tnaC, tnaA and tnaB. The upstream region of tnaA is a transcribed regulatory region (tnaL) that contains a small translated regulatory protein tnaC

Figure 13 shows the identification of a transcribed region (indicated with the arrow). The region was previously not known as a transcribed region.

**Conclusion**

The present invention provides greatly improved methods for transcriptional annotation. It is to be understood that the above description is intended to be illustrative and not restrictive. Many variations of the invention will be apparent to those of skill in the art upon reviewing the above description. By way of example, the invention has been described primarily with reference to the use of a high density oligonucleotide array, but it will be readily recognized by those of skill in the art that other nucleic acid arrays, other methods of measuring transcript levels and gene expression monitoring at the protein level could be used.